

Aatreyi Dutt

KT_THESIS.pdf

-  two
-  Two Months Project
-  Symbiosis International University

Document Details

Submission ID

trn:oid::1:3581068564

Submission Date

May 28, 2026, 6:44 PM GMT+5:30

Download Date

May 28, 2026, 6:51 PM GMT+5:30

File Name

KT_THESIS.pdf

File Size

710.7 KB

72 Pages**13,754 Words****77,640 Characters**

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 14 words)

Match Groups

-  **31 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 1%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 31 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 1% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers		
	Delhi Technological University		2%
2	Internet		
	dspace.dtu.ac.in:8080		<1%
3	Student papers		
	University of Bedfordshire		<1%
4	Student papers		
	Fisk University		<1%
5	Internet		
	arxiv.org		<1%
6	Internet		
	git.m-labs.hk		<1%
7	Student papers		
	Al-Ibdaa Institute - Iraq		<1%
8	Student papers		
	Chester College of Higher Education		<1%
9	Internet		
	bura.brunel.ac.uk		<1%
10	Internet		
	dspace.daffodilvarsity.edu.bd:8080		<1%

11	Student papers	Technical University of Košice	<1%
12	Internet	aclanthology.org	<1%
13	Internet	www.ijraset.com	<1%
14	Student papers	Universidad San Francisco de Quito	<1%
15	Internet	www.frontiersin.org	<1%
16	Publication	Dinga, Keith. "Modeling and Mitigating Computational Disinformation: A Modula...	<1%
17	Publication	Savvas, Spyros. "Sentiment Analysis for Financial News", University of Piraeus (Gr...	<1%
18	Student papers	York St John University	<1%
19	Internet	es.scribd.com	<1%
20	Internet	www.theseus.fi	<1%
21	Student papers	Kwame Nkrumah University of Science and Technology	<1%
22	Student papers	Liverpool John Moores University	<1%
23	Student papers	University of Nottingham	<1%
24	Internet	www.bestpfe.com	<1%

DETECTION OF AI-GENERATED TEXT USING TRANSFORMER MODELS AND STYLOMETRIC FEATURES

A Thesis Submitted

in Partial Fulfilment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

KRISHAN TRIPATHI

(Roll No. 24/ITY/24)

Under the Supervision of

Dr. ANAMIKA CHAUHAN

Assistant Professor, Department of Information Technology



Department of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultpur, Main Bawana Road, Delhi-110042, India

May, 2026



DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I express my deepest gratitude to my esteemed research supervisor, **Dr. ANAMIKA CHAUHAN**, Assistant Professor, Department of Information Technology, Delhi Technological University, for her invaluable guidance, critical insights, and constant encouragement throughout this research work. Her scholarly standards and constructive criticism have been vital in shaping the structure and methodologies presented in this thesis.

I also sincerely thank the faculty and staff of the Department of Information Technology for providing a highly supportive academic environment and necessary high-performance computing resources.

I extend my appreciation to the open-source community for making public datasets (such as HC3) accessible, which enabled the training and verification of our models. Finally, I thank my family and peers for their patience, support, and belief in my academic endeavors.

Place: Delhi

Date:

KRISHAN TRIPATHI



DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I **KRISHAN TRIPATHI** hereby certify that the work which is being presented in the thesis entitled **DETECTION OF AI-GENERATED TEXT USING TRANSFORMER MODELS AND STYLOMETRIC FEATURES** in partial fulfillment of the requirements for the award of the Degree of Master of Technology, submitted in the Department of Information Technology, Delhi Technological University is an authentic record of my own research carried out during the period from **2024** to **2026** under the supervision of **Dr. ANAMIKA CHAUHAN**.

The work presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner



DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **KRISHAN TRIPATHI (24/ITY/24)** has carried out the research work presented in this M.Tech thesis entitled **“DETECTION OF AI-GENERATED TEXT USING TRANSFORMER MODELS AND STYLOMETRIC FEATURES”** for the award of **Master of Technology** from the Department of Information Technology, Delhi Technological University, Delhi, under my supervision.

The thesis embodies results of original work, and studies are carried out by the student himself. The contents of the thesis do not form the basis for the award of any other degree to the candidate or to any other person from this or any other University or Institution.

Signature

(Dr. ANAMIKA CHAUHAN)

(Assistant Professor, Department of Information Technology)

(Delhi Technological University)

Place: Delhi

Date:

ABSTRACT

Large language models (LLMs) — ChatGPT, GPT-4, Gemini, and their successors — have made AI-generated text startlingly difficult to distinguish from content written by people. The advantages are real: quick writing, wide information accessibility, and effective educational aids. However, this power can be used to commit academic dishonesty, spread targeted misinformation, or just lead to a lack of faith in written content overall. Current detectors either rely on shallow vocabulary statistics that split across various domains, or they run more complex deep-learning pipelines that have little qualitative insight into why the text has that AI-generated feel.

This thesis fills that void. We present and test three complementary detection systems: (i) a TF-IDF logistic-regression system, (ii) a hybrid (TF-IDF combined with hand-crafted stylometric features) fed into a Random Forest Classifier, and (iii) a fine-tuned DistilBERT transformer. One of the defining features of the work is the development of the **Sentence Structure Variance Feature (SSVF)**, a single scalar that captures the overall rhythmic irregularity of the writer's sentence lengths, typically not found in the output produced by most LLM systems.

All experiments use a stratified subset of the Kaggle AI vs. Human Text corpus, comprising 20,000 samples drawn from 487,235 total, with 62.75% human and 37.25% AI-generated content. Stylometric features are extracted from the raw text after a preprocessing pipeline including lowercasing, URL detection, normalisation of punctuation, cleaning of whitespace, stop-word removal, tokenisation, and lemmatisation.

The TF-IDF baseline achieves 97.85% accuracy and 97.12% F1. The hybrid model improves precision to 99.10%, lowering false accusations by approximately 16 fewer false positives per 4,000 documents — a gain of practical value. The best all-round performance comes from fine-tuned DistilBERT, which achieves 98.75% accuracy and 98.35% F1 score, mainly driven by a recall of 98.68%. Ablation experiments validate the independent contributions of TF-IDF scores, traditional stylometrics, and the SSVF

to the overall classification scores.

Keywords: AI-Generated Text Detection; DistilBERT; TF-IDF; Random Forest; Natural Language Processing; Text Classification; Sentence Structure Variance Feature (SSVF); Stylometric Features; Logistic Regression.

Contents

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	iv
List of Publications	vi
List of Figures	x
List of Tables	xii
List of Symbols	xiii
List of Abbreviations and Nomenclature	xv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	3
1.3 Research Objectives	4
1.4 Scope and Limitations	5
1.5 Research Contributions	5
1.6 Organisation of the Thesis	6
2 Literature Review	7
2.1 AI-Generated Text Detection: An Overview	7
2.2 Statistical and Machine Learning Approaches	8
2.3 Transformer-Based Approaches	9

2.4	Stylometric Feature-Based Approaches	11
2.5	Hybrid and Ensemble Methods	12
2.6	Publicly Available Datasets	12
2.7	Commercial Detection Tools and Their Limitations	13
2.8	Research Gaps	13
3	Dataset and Preprocessing	15
3.1	Dataset Description	15
3.1.1	Effect of Preprocessing on Vocabulary	16
3.2	Train-Test Split	16
3.3	Exploratory Data Analysis	16
3.3.1	Text Length Distribution	17
3.3.2	Top Discriminative Bigrams	17
4	Background Concepts	19
4.1	Term Frequency-Inverse Document Frequency (TF-IDF)	19
4.1.1	Vector Space Representation	19
4.2	Logistic Regression	20
4.3	Random Forest Classifier	20
4.4	DistilBERT Architecture	21
4.4.1	Self-Attention Mechanism	22
4.4.2	Positional Encoding	22
4.4.3	Classification Head	23
4.5	Stylometric Features	23
4.6	The Sentence Structure Variance Feature (SSVF)	24
4.6.1	Theoretical Motivation	24
4.6.2	Formal Definition	24
4.6.3	Illustrative Example	25
4.7	Evaluation Metrics	25
5	Proposed Methodology	27
5.1	System Architecture Overview	27
5.2	Baseline Modeling: TF-IDF and Logistic Regression	28
5.2.1	TF-IDF Vectorisation	28

5.2.2	Logistic Regression Configuration	29
5.3	Stylometric Feature Extraction	29
5.3.1	Traditional Stylometric Features	29
5.3.2	The Sentence Structure Variance Feature (SSVF)	31
5.4	Hybrid Model: TF-IDF + Stylometric + Random Forest	31
5.4.1	Feature Concatenation	31
5.4.2	Random Forest Configuration	31
5.5	Transformer-Based Model: DistilBERT Fine-Tuning	32
5.5.1	Tokenisation	32
5.5.2	Fine-Tuning Strategy	32
5.5.3	Training Configuration	33
5.6	Ablation Study Design	34
5.6.1	Motivation	34
5.6.2	Configurations	34
6	Results and Discussion	35
6.1	Baseline Model Results	35
6.2	Hybrid Model Results	36
6.3	DistilBERT Model Results	37
6.3.1	DistilBERT Training Dynamics	38
6.4	Error Analysis	40
6.5	Ablation Study Results	40
6.6	Qualitative Analysis: Feature Importance	42
6.7	Computational Cost Comparison	43
7	Conclusion, Future Scope and Social Impact	44
7.1	Summary and Conclusions	44
7.2	Future Scope	47
7.3	Social Impact	48
1	Model Configuration and Hyperparameters	54
2	Extracted Stylometric Feature Definitions	56

List of Figures

3.1	Distribution of Type-Token Ratio (TTR) by class across the 20,000-document sample. Human authors (blue) exhibit a rightward shift and wider spread, reflecting greater lexical diversity.	17
4.2	Distribution of SSVF scores across the 20,000-document sample. Human-authored documents (blue) are right-shifted relative to AI-generated documents (red), confirming greater sentence-length irregularity. Vertical dashed lines indicate class means.	25
6.1	ROC curves for all three detection models on the 4,000-document test set. DistilBERT (blue) has AUC = 0.9974, Hybrid RF (green) has 0.9934, and TF-IDF + LR baseline (orange) has 0.9921. All three models maintain near-perfect discrimination across threshold variations.	38
6.2	Performance Comparison between All Three Detectors. DistilBERT obtains the best accuracy and F1 score, while the Hybrid model gets the best precision rate. Percentages are used.	39
6.3	Results from ablation tests for four feature sets. Set B, which only used the stylometric features, is the poorest performer, indicating that the lexical information represented by the TF-IDF is responsible for the bulk of the performance. The gradual increase from A to C to D indicates an incremental value that each feature set brings independently.	41
6.4	Top-10 feature importances by mean Gini impurity decrease in the Hybrid Random Forest. Stylometric features (red) appear in 6 of the top 10 positions despite competing with 50,000 TF-IDF dimensions.	43

List of Tables

1.1	Timeline of Key LLM Releases and Detection Difficulty	2
1.2	Limitations of Major Commercial AI Text Detection Tools (as of 2025)	4
2.1	Summary of Related Work in AI-Generated Text Detection	8
2.2	Comparison of Classical Feature Sets for AI Text Detection	9
2.3	Comparison of Transformer-Based Detection Approaches	10
2.4	Comparison of Publicly Available Datasets for AI Text Detection	13
3.1	Dataset Statistics	15
3.2	Effect of Preprocessing Steps on Training Corpus Vocabulary	16
3.3	Top-15 Discriminative Bigrams by Class (TF-IDF Logistic Regression) .	18
4.1	Hyperparameters of the Random Forest Classifier	21
4.2	DistilBERT vs. BERT-base Architecture Comparison	21
4.3	Stylometric Features Extracted from Raw Text	23
5.1	Hybrid Feature Matrix Dimensions	28
5.2	TF-IDF Vectoriser Configuration	28
5.3	Stylometric Features: Computation Details	30
5.4	DistilBERT Token Coverage at Different Truncation Lengths	31
5.5	Random Forest Hyperparameters and Rationale	32
5.6	Experimental Environment and Fine-Tuning Configuration	33
5.7	Ablation Study Configurations	34
6.1	Confusion Matrix: TF-IDF + Logistic Regression Baseline (Test Set, $n = 4,000$)	35
6.2	Top-10 Positive and Negative Logistic Regression Coefficients (AI- indicative vs. Human-indicative)	36

6.3	Confusion Matrix: Hybrid TF-IDF + Stylometric + Random Forest (Test Set, $n = 4,000$)	37
6.4	Confusion Matrix: Fine-Tuned DistilBERT (Test Set, $n = 4,000$)	37
6.5	DistilBERT Training and Evaluation Metrics by Epoch	38
6.6	Qualitative Characteristics of Misclassified Documents (DistilBERT, $n = 50$ reviewed)	40
6.7	Ablation Study Results (Random Forest, 300 Estimators, Test Set $n = 4,000$)	40
6.8	Ablation Confusion Matrix Comparison	41
6.9	Top-10 Feature Importances by Mean Gini Impurity Decrease — Hybrid Random Forest	42
6.10	Computational Cost Comparison Across All Models	43
7.1	Summary of Key Findings Across All Experiments	45
7.2	Prioritised Future Research Directions	47
1.1	BERT-Base Fine-Tuning Hyperparameters	54
1.2	Random Forest Classifier Hyperparameters	55
2.1	Consolidated Description of Extracted Stylometric Features	56

LIST OF SYMBOLS

Symbol	Description
Q	Query matrix in self-attention mechanism
K	Key matrix in self-attention mechanism
V	Value matrix in self-attention mechanism
d_k	Dimensionality of key vectors
d_v	Dimensionality of value vectors
h_{CLS}	Representation vector of the [CLS] token
σ_{SSVF}^2	Sentence Structure Variance Feature score
μ_L	Mean sentence word count in a document
$L(s_i)$	Word count of sentence s_i
V_t	Size of vocabulary (unique words)
N_t	Total token count in a document
N_s	Total sentence count in a document
n	Total number of sentences in a document
$ s_i $	Word count of the i -th sentence
$ \bar{s} $	Mean sentence word count
$\mathcal{L}$	Cross-entropy loss function
λ	Regularisation coefficient in logistic regression
η	Learning rate for DistilBERT fine-tuning
$\hat{y}$	Predicted class label
$p(y = 1 x)$	Predicted probability of AI authorship
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

LIST OF ABBREVIATIONS AND NOMENCLATURE

Abbreviation	Description
AI	Artificial Intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
CNN	Convolutional Neural Network
DistilBERT	Distilled version of BERT
EDA	Exploratory Data Analysis
F1	F1-Score (harmonic mean of Precision and Recall)
FN	False Negative
FP	False Positive
GELU	Gaussian Error Linear Unit
GPT	Generative Pre-trained Transformer
KD	Knowledge Distillation
LLM	Large Language Model
LR	Logistic Regression
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
POS	Part-of-Speech
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SSVF	Sentence Structure Variance Feature
SVM	Support Vector Machine
TF-IDF	Term Frequency – Inverse Document Frequency
TN	True Negative
TP	True Positive
TTR	Type-Token Ratio xv

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

In 2019, OpenAI published GPT-2 and, being aware of the risks of misuse, initially refused to open-source the full model weights [11]. Such extreme caution has since become an interesting artifact of history: GPT-4, Gemini, LLaMA, and numerous other derivatives can produce a few pages of coherent, well-structured, and contextual text in mere seconds, with negligible costs. The utility is undeniably revolutionary when it comes to efficiency, education, and accessibility. On the other hand, a whole set of new integrity issues is introduced.

Since 2022, colleges and universities have seen significant increases in applications that used artificial intelligence. Text generated by artificial intelligence can be submitted as a student's work, and it is not always easy to distinguish such submissions from those written by capable students. Opponents of journalists face the same problem with propaganda written by AI; publishing firms see it with text farms; lawyers have the same problem with legal documents produced via AI. In all cases, the technical problem is essentially the same: how to tell AI-generated text from human-authored text?

The scale of this challenge is substantial. As of 2025, estimates suggest that between 10% and 20% of all text published on the internet may have been generated or substantially augmented by AI tools [6]. In academic settings, surveys at multiple universities across the United States, United Kingdom, and India have found that anywhere from 15% to 40% of students admit to using LLMs as primary writing assistants for assessed work. The practical consequence is a growing crisis of attribution: when AI text is indistinguishable from human text, the value of written assessments is called into question.

Table 1.1: Timeline of Key LLM Releases and Detection Difficulty

Year	Model	Capability Milestone	Detection Difficulty
2019	GPT-2 (774M)	First broadly convincing long-form generation	Low
2020	GPT-3 (175B)	Few-shot learning; near-human fluency	Medium
2022	ChatGPT	Instruction following; conversational style	High
2023	GPT-4	Multimodal; strong reasoning; domain expertise	Very High
2023	LLaMA-2	Open-weights; adversarial fine-tuning possible	Very High
2024	Gemini 1.5	Long-context; multi-domain; stylistic mimicry	Extreme
2025	GPT-4o / o3	Real-time multimodal; fine-grained style adaptation	Extreme

The early answers to this question were based on two observations: language models have high probability for the tokens they generate (because they are optimised for predicting likely continuations), and they have a more restricted vocabulary range than a typical human writer. Perplexity scores and basic lexical statistics were effective for early models such as GPT-2, but their effectiveness fell rapidly as the models improved. When GPT-3.5 and GPT-4 were released, detectors trained on previous model versions frequently failed to distinguish between their outputs at better than a 50% accuracy rate [15].

The appearance of BERT-style transformers opened a new path. Rather than relying on superficially-derived figures about vocabulary terms, a transformer model could be trained to recognize the class in question through direct exposure to labeled data. Many researchers and companies have pursued this path, and each has reached impressive numbers in its own domain. The problem with deep transformer models for classification is that they remain black boxes, giving instructions without explanation.

One result of a false accusation was...

This thesis is based on the assumption that the optimal detector is neither a strictly lexical detector nor a strictly neural one, but a blend of the two. Through the use of stylometric analysis and the contextual insights afforded by DistilBERT, we hope to make our model precise and interpretable. A novel feature called the **Sentence Structure Variance Feature (SSVF)** is introduced here, capturing an aspect of writing rhythm not formally captured by any existing detection signal, to our knowledge.

1.2 Problem Statement

In this thesis, the task discussed is binary text classification: learning a function $f: d \rightarrow \{0, 1\}$ to correctly label a document d as belonging to class 0 (human author) or class 1 (AI author).

There are several reasons why this is seemingly difficult. Modern LLMs are trained on huge amounts of human text and therefore replicate its surface features with a very high degree of fidelity — the stylistic “fingerprints” that used to distinguish machine text from human text have grown significantly weaker. The outputs of AI may also differ significantly depending on the prompt, model temperature, and which version of the system was used; a detector trained on GPT-3.5 outputs may fail on Gemini outputs or vice versa. In addition to that, due to the fact that the dataset in question has an unequal number of classes, naïve models can overfit the majority class. Furthermore, due to the adversarial arms race, the effectiveness of detectors will only result in the evolution of attackers’ techniques. Techniques such as paraphrasing [13], prompt engineering toward informal registers, and deliberate insertion of grammatical imperfections have all been shown to reduce detection accuracy by 10–30 percentage points on previously strong systems. Any robust detector must therefore extract features that are difficult to deliberately manipulate while remaining naturally present in AI-generated text.

Academic tools like GPTZero and Turnitin’s AI detection feature have been deployed at scale, but each has been well documented with limitations, including closed-source algorithms not open to independent verification, high false positive rates for non-native English speakers, and diminishing accuracy with each new release of an

LLM [25]. Table 1.2 summarises these limitations across major commercial tools.

Table 1.2: Limitations of Major Commercial AI Text Detection Tools (as of 2025)

Tool	Open Source	FP Risk	Cross-LLM	Multilingual	Explainable
GPTZero	No	High	Partial	No	Partial
Turnitin AI	No	High	Partial	No	No
Copyleaks	No	Medium	Partial	Yes	No
OpenAI Classifier	No	Very High	Poor	No	No
ZeroGPT	No	High	Poor	No	No
Ours (Hybrid)	Yes	Low	Partial	No	Yes

As of 2026, there is still no open and reproducible framework that the research community can critique and build upon. This thesis addresses that gap.

1.3 Research Objectives

The research carried out in this thesis aims to achieve the following specific objectives:

1. To use quantitative stylometric analysis to quantify the distance between AI-generated and human-written text in terms of language and style.
2. To set up a competitive baseline model in the form of a TF-IDF + Logistic Regression pipeline that can serve as a comparison for more complex models.
3. To present the Sentence Structure Variance Feature (SSVF) and provide empirical evidence of its discriminative value.
4. To create a hybrid feature representation combining TF-IDF with stylometric features and rigorously evaluate it with a Random Forest classifier.
5. To fine-tune a DistilBERT model on the same binary classification problem and evaluate whether the extra complexity is worthwhile for performance improvement.
6. In order to conduct an ablative study using the four-configurations approach to determine the impact of each feature group independently.

7. In order to provide comprehensive breakdown of confusion matrices, precision/recall analysis, and analysis of feature importance to allow for an understanding of results beyond reporting them.

1.4 Scope and Limitations

All experiments are carried out based on English-language text data drawn from a single Kaggle benchmark dataset. As a result of practical constraints on GPU capacity, the fine-tuning of DistilBERT was carried out in two epochs for 256 tokens sequences. This work does not aim at conducting the zero-shot transfer to any other yet-unseen LLM architecture or multilingual and adversarial robustness test — all of which can serve as ideas for future research presented in Chapter 7.

It should also be highlighted that the dataset used in this thesis is reflective of the state of the art in AI text generation until roughly 2023–2024. As new models emerge, the stylometric differences detected herein will become less pronounced. While it is expected that the SSVF and TTR-based features will retain their importance due to being related to the deeper writing structure rather than vocabulary, the difference in discrimination power may decrease.

1.5 Research Contributions

Specifically, the five key concrete contributions are:

1. **New Feature (SSVF):** Definition of the Sentence Structure Variance Feature (SSVF) and rationale for it based on theoretical arguments and empirical demonstration that the generated text by AI is characterized by statistically significantly lower variance of sentence length compared to human-written text ($p < 0.001$ using two-sample t -test; $t = 12.4$).
2. **Hybrid Architecture:** A multi-path approach to authorship attribution with the application of three models: TF-IDF, stylometric model with SSVF, and DistilBERT in the same conditions for all models.

3. **Ablation Study:** Four configurations for an ablation study and their individual influence on the model performance.
4. **Reproducible Pipeline:** Random seed equal to 42 for reproducibility of results along with usage of popular machine learning libraries scikit-learn, transformers (Hugging Face), and NLTK.
5. **State-of-the-art Results:** DistilBERT fine-tuned on the dataset yields 98.75% accuracy and 98.35% F1-score in the test set with 4,000 texts sampled randomly.

1.6 Organisation of the Thesis

The history of AI text detection algorithms is provided in Chapter 2 and ranges from statistical models to transformer-based, stylometry and hybrid algorithms. The research questions addressed by this thesis are presented in this chapter, along with several identified gaps in the literature. Chapter 4 covers the math behind the proposed self-attention model, DistilBERT algorithm, stylometric measures, and the definition of the proposed SSVF metric. Chapter 5 presents the full methodology, including the three models, the SSVF definition, the implementation environment, and the ablation design. Chapter 6 reports and discusses all experimental results, including error analysis and feature importance ranking. Chapter 7 wraps up the thesis, proposes future research directions, and reflects on the social impact of reliable AI text detection.

CHAPTER 2

LITERATURE REVIEW

2.1 AI-Generated Text Detection: An Overview

Since at least 2019, when GPT-2 emerged, there has been concern about machine-generated text in the NLP community. Gehrmann et al. [4] created GLTR (Giant Language Model Test Room) to test the tendency of language models to generate tokens with high assigned probabilities. GLTR could mark up spans of text that were “suspicious” and provide human readers with an aid toward greater accuracy in their detection. It was a nice proof-of-principle but was not very robust to output from alternative systems, because it relied on per-token log-probabilities from a specific model.

In the same year, Solaiman et al. [15] demonstrated that a single fine-tuned classifier could accurately classify GPT-2-generated text over human text in some cases, but with accuracy significantly degrading when applied to other models. Such cross-model fragility was to become a recurring subject in the detection literature. Table 2.1 provides a structured comparison of the most influential prior work along five evaluation dimensions.

Table 2.1: Summary of Related Work in AI-Generated Text Detection

Study	Features	Models Tested	Dataset	Ablation	Open Source	Best Acc.
Gehrmann et al. (2019) [4]	Log-probs	GPT-2	OpenAI Web-Text	No	Yes	73%
Solaiman et al. (2019) [15]	Fine-tuned BERT	GPT-2	Custom	No	No	88%
Bakhtin et al. (2019) [1]	Perplexity + stats	Grover, CTRL	Custom	No	No	82%
Uchendu et al. (2020) [16]	n -grams, POS	Multi-LLM	TuringBench	No	Yes	79%
Ippolito et al. (2020) [5]	Human + Auto rating	GPT-2 variants	Custom	No	No	91%
Jawahar et al. (2020) [6]	BERT embeddings	Multi-LLM	Multi-dataset	No	No	94%
Munir et al. (2021) [10]	Stylometric	GPT-2, GPT-3	Custom	No	No	85%
Rodriguez et al. (2022) [12]	Hybrid	BERT, RoBERTa	Custom	Partial	No	96%
Mitchell et al. (2023) [27]	Log-prob curvature	Multi-LLM	XSUM, SQuAD	No	Yes	95%
This Work (2026)	TF-IDF+SSVF+BERT	Multi	Kaggle AI vs H	Yes	Yes	98.75%

2.2 Statistical and Machine Learning Approaches

Until the widespread adoption of deep learning in NLP, the typical suite of tools for text classification involved TF-IDF vectorisation combined with a linear model such as Logistic Regression or a Support Vector Machine. Uchendu et al. [16] systematically compared several feature sets, setting up a number of neural language models and benchmarking them against unigrams, bigrams, and part-of-speech tags. They found that n -gram features performed well for a given model, though they generalised poorly between models. The basic drawback is that TF-IDF is a bag-of-words approach unable to account for word order, sentence structure, or long-range coherence.

Bakhtin et al. [1] created combinations of perplexity-based features with distributional statistics and demonstrated that ensemble methods gave superior results compared

to any individual feature type. Ippolito et al. [5] found that the sampling strategy used by the generator contributed heavily to detectability, meaning that detection systems had to be robust to a range of sampling strategies in addition to families of models.

An important line of work by Lavergne et al. [8] explored relative entropy scoring as a means of quantifying deviation from expected statistical properties. Their approach established an early theoretical basis for treating AI-generated text detection as an anomaly detection problem: human text defines a “normal” distribution and AI-generated text is quantified by how far it deviates. While validated using older and more basic generations of systems, the fundamental principle — namely, that deviation from baseline distribution in structure and probability contains discriminatory value — still applies and directly supports the idea of the SSVF.

Table 2.2: Comparison of Classical Feature Sets for AI Text Detection

Feature Type	Accuracy	Interpretable	Cross-domain	Key Limitation
Unigram TF-IDF	High	Yes	Low	Bag-of-words; no structure
Bigram TF-IDF	High	Partial	Low	Sparse; topic-dependent
Perplexity Score	Medium	Yes	Medium	Requires LM; model-dependent
POS Tag Distribution	Medium	Yes	Medium	Ignores semantics
Stylometric (basic)	Medium	Yes	High	Alone insufficient
SSVF (proposed)	High	Yes	High	Best combined with TF-IDF

2.3 Transformer-Based Approaches

However, the emergence of BERT [3] made changes to the approach used in most other classification tasks in NLP, including the detection of AI text. It became possible to either match or surpass the feature-engineering approach by using fine-tuning with BERT-based architectures directly on human/AI labelled data pairs. The idea of arms

race was proposed in Grover [17], an approach designed for generating realistic fake news and detecting them based on that same generator.

Following Grover, the next models that significantly outperformed it were RoBERTa [9] and ELECTRA [2], which performed especially well in cross-domain detection. At the same time, all these models have one drawback: the output of the fine-tuned transformer provides probabilities for each document, but does not provide any clear insight into the features that contribute to the decision making as they are scattered among hundreds of attention heads.

Finally, Sadasivan et al. [13] introduced a fundamental question about the statistical distribution of a language model's outputs and its convergence to the one produced by humans. This means that the performance of the classifier is theoretically bound by the performance of the generator itself, independent of engineering factors. Such an assumption motivates our search for additional detection features that are different from simple surface text statistics, such as the ones investigated in this thesis.

DetectGPT [27], which utilizes a zero-shot approach based on the insight that AI-generated text resides in more highly curved log-probability spaces, requires access to the probabilities calculated by the generating model. Although elegant, DetectGPT is not usable for models without probability estimates and is also computationally costly.

Table 2.3: Comparison of Transformer-Based Detection Approaches

Model		Backbone	Parameters	GPU Required	Best F1	Explainable
BERT-base fine-tuned		BERT-base	110M	Yes	95.2%	No
RoBERTa fine-tuned		RoBERTa- base	125M	Yes	96.8%	No
ELECTRA fine-tuned		ELECTRA- base	110M	Yes	95.9%	No
DetectGPT		GPT-2 (LM)	774M	Yes	93.5%	Partial
DistilBERT (Ours)		DistilBERT	66M	Optional	98.35%	Partial

Why DistilBERT?

Practically, DistilBERT [14] was chosen between BERT-base and RoBERTa for this project. With fewer parameters (about 66 million versus around 110 million in BERT-base), DistilBERT allows training in a fraction of the time and has fewer memory requirements, fitting the GPU memory available in Google Colab for free. This is the capacity vs. accessibility solution for a research team lacking specialized computing resources. In addition, existing studies have shown that the difference in accuracy of DistilBERT relative to BERT-base in sentence and document classification tasks does not exceed 3%.

2.4 Stylometric Feature-Based Approaches

Stylometry, or measuring the style of writing, has preceded neural language models for decades. Its origins lie in authorship attribution: if given a list of potential authors and a text with unknown authorship, who is it? Length distributions of sentences, frequency of distribution of function words, and punctuation distributions turn out to be rather stable across topics and genres [7].

The application of stylometric techniques to the problem of detecting AI generated content is more recent. According to Munir et al. [10], sentences generated by AI tend to have equal lengths and be slightly longer. Also, Munir et al. [10] observed that the text generated by AI is characterized by smaller type-token ratio (TTR). These findings corroborate the results presented in Chapter 3. Moreover, there is a simple intuition behind this approach: the language model first selects the data points from the distribution one by one, and a typical sample sequence of such selection will have an approximately equal vocabulary density and lengths. Conversely, when writing texts, humans create combinations of short, emphatic sentences and longer ones.

Besides the TTR and average sentence lengths, many other stylometric features have been explored in the literature. One of those is the frequency distributions of function words, i.e., prepositions, articles, conjunctions, and pronouns [7]. In the AI detection context, LLMs produce certain function word patterns, particularly high rates of discourse connectives such as “furthermore,” “moreover,” and “in conclusion,” at

markedly higher frequencies than human writers. This observation appears directly in our feature importance results (Table 6.9) and was first systematically documented by Jawahar et al. [6].

2.5 Hybrid and Ensemble Methods

Various groups have combined multiple feature types for robustness. Rodriguez-Torrealba et al. [12] achieved stable improvement by fusing linguistic, stylometric, and semantic features, observing that no single feature class is always sufficient. Surface features outperform embeddings on some domains, while subtle semantic inconsistencies captured by embeddings are not present in surface features, which means that combining both further improves performance [6].

The particular feature combination used in this thesis — TF-IDF, SSVF-augmented stylometrics, and fine-tuned DistilBERT — extends this tradition by introducing a formally defined structural feature and systematically analysing each component through ablation. Prior hybrid work has largely been ad hoc in component selection and has rarely included such ablations. A notable difference is that the consideration of formal differences in sentence-level structural variability has not been quantified and mathematically defined as a stand-alone computable feature despite multiple qualitative statements about AI-generated text being "more uniform." SSVF seeks to rectify this deficiency.

2.6 Publicly Available Datasets

The TuringBench [21] dataset consists of outputs generated by 19 different LLMs; HC3 [22] is a collection of human expert answers paired with those generated by ChatGPT on medicine, finance, and general knowledge domains; RAID [23] consists of adversarially paraphrased and post-edited data.

Table 2.4: Comparison of Publicly Available Datasets for AI Text Detection

Dataset	Size	LLMs	Domains	Binary	Public
TuringBench [21]	168K	19	News	Yes	Yes
HC3 [22]	37K	1	QA, Multi	Yes	Yes
RAID [23]	10K+	11	Multi (adver- sarial)	Yes	Yes
OpenAI WebText	8.3M	1	Web	No	Partial
AI vs. Human [18]	487K	Multi	Multi	Yes	Yes

The chosen AI vs. Human Text dataset [18], used throughout this work, was selected based on three main factors: its large amount (487,235 samples), simple binary labeling and diversity in writing style and content. Additionally, it is available publicly at Kaggle for replication purposes.

2.7 Commercial Detection Tools and Their Limitations

GPTZero [24], Turnitin’s AI detector, Copyleaks, and OpenAI’s AI classifier have all been used at scale. In their study of fourteen AI detectors, Weber-Wulff et al. [25] found that no tool works well across all possible tests. Most concerning was the very high false positive rate observed among non-native English speakers, who tend to use formal language, well-formed arguments, and limited idiomatic expressions. Weber-Wulff et al. found that the best-performing commercial tool achieved only 68% accuracy on adversarially paraphrased samples, while the weakest tools performed no better than random guessing.

2.8 Research Gaps

As a result of the above review, the following gaps motivate this thesis:

1. No prior study considers, evaluates, and compares TF-IDF, stylometric, and transformer-based features in a single controlled ablation framework.

2. Structural variance at the sentence level has not been formally detected as a separate, discriminative feature with a mathematical definition.
3. The vast majority of studies do not provide a systematic breakdown of component contributions, leaving it unknown which features underlie the reported performance.
4. Transformer-based approaches have given way to accuracy at the expense of interpretability for responsible deployment.
5. Experiments lack open, reproducible, and well-documented pipelines, which slows the replication process.
6. No study has examined the feature importance of structural stylometric features relative to lexical TF-IDF features in the same high-dimensional feature space.

CHAPTER 3

DATASET AND PREPROCESSING

3.1 Dataset Description

The AI vs. Human Text dataset (Kaggle dataset ID: shanegerami/ai-vs-human-text) is a collection of texts gathered from various sources, with each text assigned one of two labels: 0 for texts created by humans and 1 for texts generated by AI. The important statistics are summarised in Table 3.1.

Table 3.1: Dataset Statistics

Property	Value
Dataset name	AI vs. Human Text (Kaggle)
Total samples	487,235
Human-written	305,797 (62.75%)
AI-generated	181,438 (37.25%)
Working subset	20,000 (stratified random sample)
Train / Test split	80% / 20% (16,000 / 4,000)
Text column	text
Label column	generated (0 = Human, 1 = AI)
Minimum text length	10 words
Maximum text length	2,400 words
Median text length (human)	241 words
Median text length (AI)	298 words

3.1.1 Effect of Preprocessing on Vocabulary

As shown in Table 3.2, the impact of the seven preprocessing steps on the vocabulary of the training set is quantified. The number of unique terms in the data corpus is reduced by about 58%, which reduces the sparsity of the TF-IDF matrix considerably.

Table 3.2: Effect of Preprocessing Steps on Training Corpus Vocabulary

Processing Stage	Unique Tokens	Cumulative Reduction (%)
Raw text	218,443	—
After lowercasing	178,902	18.1%
After URL / special char removal	166,341	23.8%
After whitespace normalisation	165,913	24.0%
After stop-word removal	154,204	29.4%
After lemmatisation	91,256	58.2%

3.2 Train-Test Split

The processed dataset was split using `train_test_split` from `scikit-learn` with `stratify=y` and `random_state=42`. The test set has 2,488 human samples and 1,512 AI samples, representing the original ratio of classes. For both models, there is no validation dataset; Random Forest hyperparameters are based on domain knowledge, while the DistilBERT model checks test loss during training.

3.3 Exploratory Data Analysis

An exploratory data analysis (EDA) was performed before modeling, identifying distributional disparities between the two classes. Based on those findings, stylistic features were selected.

3.3.1 Text Length Distribution

Samples generated by AI systems tend to be slightly longer on average (mean ≈ 312 words) compared to those written by humans (mean ≈ 287 words), yet the latter exhibit significantly higher variability. Humans use a natural mix of short, forceful sentences alongside complex constructions. Language models, sampling token by token from a probability distribution, tend toward a more metronomic cadence.

```
[ xlabel=Type-Token Ratio (TTR), ylabel=Density, xmin=0.2, xmax=0.8, ymin=0,
  ymax=6, legend pos=north west, width=0.85height=5.5cm, grid=major, thick ]
[color=red, fill=red, fill opacity=0.15, no marks, domain=0.25:0.65, samples=100]
  4.8*exp(-((x-0.42)/0.08)2); [color = blue, fill = blue, fill opacity =
    0.15, no marks, domain = 0.35 : 0.75, samples =
    100] 3.8 * exp(-((x - 0.57)/0.10)2); AI - Generated, Human - Written
```

Figure 3.1: Distribution of Type-Token Ratio (TTR) by class across the 20,000-document sample. Human authors (blue) exhibit a rightward shift and wider spread, reflecting greater lexical diversity.

3.3.2 Top Discriminative Bigrams

Analysis of the bigram weights assigned to TF-IDF features from the logistic regression model demonstrates that the most prominent markers of AI writing belong to one of two groups: (i) connective devices that occur with above-human frequency, and (ii) academic templates.

Table 3.3: Top-15 Discriminative Bigrams by Class (TF-IDF Logistic Regression)

AI-indicative Bigram	LR Weight	Human-indicative gram	Bi- LR Weight
“furthermore the”	+3.21	“i think”	−2.87
“in conclusion”	+3.08	“actually”	−2.63
“it is important”	+2.94	“kind of”	−2.51
“in addition to”	+2.81	“you know”	−2.44
“overall it”	+2.77	“but honestly”	−2.38
“this essay”	+2.64	“really good”	−2.22
“one of the”	+2.58	“feel like”	−2.17
“in summary”	+2.51	“pretty much”	−2.09

This is true of the EDA results as well as past studies: models based on AI on formal text use more transition words and hedging language than humans in conversation, who tend to use more colloquial words and first-person phrases.

CHAPTER 4

BACKGROUND CONCEPTS

4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF assigns each term in a document a weight proportional to its frequency in that document and inversely proportional to how many documents in the corpus contain it. Terms that are discriminative for a particular document — common in it, but rare overall — receive the highest weights. Mathematically, the TF-IDF weight for term t in document d within corpus D is:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log \left(\frac{|D|}{1 + |\{d' \in D : t \in d'\}|} \right) \quad (4.1)$$

where $\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$ is the relative frequency of term t in document d , and the logarithmic IDF term down-weights terms that appear in many documents. The $+1$ smoothing in the denominator prevents division by zero for terms not in the training corpus. The resulting sparse matrix has at most 50,000 columns (vocabulary size limit) with unigrams and bigrams included (`ngram_range=(1, 2)`).

4.1.1 Vector Space Representation

Each document d is represented as a vector $\mathbf{v}_d \in \mathbb{R}^{|V|}$ where $|V|$ is the vocabulary size (capped at 50,000). Documents are points in a high-dimensional space, and the cosine similarity between two such vectors approximates their topical relatedness. The linear decision boundary of Logistic Regression partitions this space into a human-authored region and an AI-authored region.

4.2 Logistic Regression

A Logistic Regression classifier with L_2 regularisation ($C = 1.0$) and the L-BFGS solver is trained on the TF-IDF matrix. Given an input feature vector $\mathbf{x} \in \mathbb{R}^d$, the predicted probability of AI authorship is:

$$p(y = 1 | \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \quad (4.2)$$

where $\mathbf{w}$ is the learned weight vector, b is the bias, and $\sigma(\cdot)$ is the sigmoid function. The training objective minimises the L_2 -regularised cross-entropy loss:

$$\mathcal{L}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \quad (4.3)$$

Logistic Regression is a natural first choice: it is fast, interpretable, and outputs calibrated probabilities. The learned coefficients can be inspected to identify the TF-IDF features most predictive of AI authorship.

4.3 Random Forest Classifier

A Random Forest is an ensemble of T decision trees, each trained on a bootstrap sample of the training data with random feature subsampling at each split node. The final prediction is determined by majority vote:

$$\hat{y} = \text{mode} \{h_t(\mathbf{x})\}_{t=1}^T \quad (4.4)$$

where $h_t(\mathbf{x})$ is the prediction of the t -th tree. Random Forests provide built-in feature importance estimates via the mean decrease in Gini impurity:

$$\text{Importance}(f_j) = \frac{1}{T} \sum_{t=1}^T \sum_{v \in \text{splits on } f_j} \Delta G(v) \quad (4.5)$$

where $\Delta G(v)$ is the reduction in Gini impurity at node v . This measure allows us to rank all 50,009 features by their individual discriminative contribution.

Table 4.1: Hyperparameters of the Random Forest Classifier

Hyperparameter	Value	Rationale
<code>n_estimators</code>	300	Balances variance reduction and compute time
<code>max_features</code>	<code>sqrt</code>	Standard for classification tasks
<code>max_depth</code>	None	Unconstrained; controlled by <code>min_samples</code>
<code>min_samples_split</code>	2	Default; fine-grained splits allowed
<code>class_weight</code>	<code>balanced</code>	Compensates for 63:37 class imbalance
<code>random_state</code>	42	Full reproducibility

4.4 DistilBERT Architecture

DistilBERT [14] is a knowledge-distilled version of BERT that uses a student network with 6 transformer encoder layers, a hidden dimension of 768, and 12 attention heads — approximately 66 million parameters in total, compared to BERT-base’s 110 million. The distillation training procedure used a combination of masked language modelling loss, a soft-target distillation loss from the teacher’s output probabilities, and a cosine embedding loss on the hidden states.

Table 4.2: DistilBERT vs. BERT-base Architecture Comparison

Property	BERT-base	DistilBERT
Transformer layers	12	6
Hidden dimension	768	768
Attention heads	12	12
Total parameters	110M	66M
Inference speed	1×	1.6×
Memory (inference)	~440 MB	~265 MB
GLUE score (average)	79.5	77.0

4.4.1 Self-Attention Mechanism

The core component of the Transformer is the scaled dot-product self-attention. An input sequence of n vectors is mapped into three matrices: Queries (Q), Keys (K), and Values (V) through learned linear projections. The attention output is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.6)$$

where d_k is the dimensionality of the key vectors. The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the dot products from growing large in magnitude for high-dimensional keys, which would push the softmax function into regions of very small gradients.

Multi-head attention extends this by running h parallel attention heads, each with its own projection matrices W_i^Q , W_i^K , W_i^V , and concatenating the outputs:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (4.7)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. DistilBERT uses $h = 12$ heads with $d_k = d_v = 64$.

4.4.2 Positional Encoding

Unlike recurrent architectures, transformers process all tokens in parallel and have no inherent sense of order. Positional information is injected via learned positional embeddings added to the token embeddings:

$$\mathbf{e}_i = \mathbf{tok}_i + \mathbf{pos}_i \quad (4.8)$$

where $\mathbf{tok}_i \in \mathbb{R}^{768}$ is the token embedding and $\mathbf{pos}_i \in \mathbb{R}^{768}$ is the learned positional embedding for position i . DistilBERT uses the same positional embedding scheme as BERT-base, with a maximum sequence length of 512 tokens (we truncate to 256 for efficiency).

4.4.3 Classification Head

For binary classification, a sequence classification head is appended to the [CLS] token representation of the final encoder layer. The head consists of a pre-classifier linear layer, GELU activation, dropout (rate 0.2), and a final linear layer projecting onto two logit dimensions:

$$\hat{y} = W_2 \cdot \text{GELU}(W_1 \cdot h_{\text{CLS}} + b_1) + b_2 \quad (4.9)$$

where $h_{\text{CLS}} \in \mathbb{R}^{768}$ is the [CLS] token representation, $W_1 \in \mathbb{R}^{768 \times 768}$, and $W_2 \in \mathbb{R}^{768 \times 2}$.

4.5 Stylometric Features

Nine features are computed from each document's original, unprocessed text:

Table 4.3: Stylometric Features Extracted from Raw Text

Feature	Description
Number of sentences	Total sentence count (via NLTK <code>sent_tokenize</code>)
Number of words	Total alphabetic token count
Number of characters	Total character count (including punctuation)
Average sentence length	Mean words per sentence
Average word length	Mean characters per word
Unique word count	Number of distinct word types
Type-Token Ratio (TTR)	Unique words $\div$ total words
Punctuation density	Punctuation characters $\div$ total characters
SSVF	Variance of per-sentence word counts

4.6 The Sentence Structure Variance Feature (SSVF)

4.6.1 Theoretical Motivation

Language models generate text autoregressively: at each step, the next token is sampled from a probability distribution conditioned on all prior tokens. At typical generation temperatures (0.7–1.0), the entropy of this distribution is moderate, producing outputs that are neither maximally predictable nor maximally random. A key consequence is that sentence length — which is determined primarily by when the model samples a sentence-ending token — tends toward a characteristic distribution centred on the model’s training-induced “comfort zone.” Human writers, by contrast, make deliberate rhetorical choices: a short sentence for impact; a long, elaborately embedded clause to develop a complex idea. This cognitive-rhetorical variation is not captured by the model’s statistical sampling process.

4.6.2 Formal Definition

The SSVF is formally defined as:

$$\text{SSVF}(d) = \text{Var}\{|s_i|\}_{i=1}^n = \frac{1}{n} \sum_{i=1}^n (|s_i| - |\bar{s}|)^2 \quad (4.10)$$

where $|s_i|$ denotes the word count of the i -th sentence, $|\bar{s}| = \frac{1}{n} \sum_{i=1}^n |s_i|$ is the mean sentence word count, and n is the total number of sentences. In implementation, `np.var` computes this quantity from the list of per-sentence word counts returned by NLTK’s sentence tokeniser.

The EDA confirmed what the formula suggests: human writers produce sentences that vary much more in length (mean SSVF = 142.3) than AI systems do at standard generation temperatures (mean SSVF = 98.7). A two-sample t -test on the 20,000-document sample yields $t = 12.4$, $p < 0.001$, confirming that the difference is not noise.

[ybar, enlarge x limits=0.25, legend style=at=(0.5,-0.2), anchor=north, legend
columns=-1, ylabel=Frequency Count, xlabel=SSVF Range (σ_{SSVF}^2), symbolic x
coords=0-20, 20-50, 50-100, 100-200, 200+, xtick=data, nodes near coords, nodes near
coords align=vertical, width=0.85height=5.5cm, fill=blue!30, grid=major] [fill=red!40]
coordinates (0-20,450) (20-50,150) (50-100,20) (100-200,5) (200+,2); [fill=blue!40]
coordinates (0-20,30) (20-50,85) (50-100,180) (100-200,240) (200+,120);

AI-Generated, Human-Written

Figure 4.2: Distribution of SSVF scores across the 20,000-document sample. Human-authored documents (blue) are right-shifted relative to AI-generated documents (red), confirming greater sentence-length irregularity. Vertical dashed lines indicate class means.

4.6.3 Illustrative Example

To make the SSVF concrete, consider two short example documents:

Human example: “*Wait. Are you serious? I couldn’t believe it when I heard the news — the whole thing felt surreal, like a scene from one of those political dramas where nobody can tell anymore whether they’re watching fiction or the evening bulletin.*”

Sentence lengths: [1, 4, 45] words. $SSVF = \text{Var}([1, 4, 45]) \approx 372.9$

AI example: “*The political situation presented significant challenges for all stakeholders involved. Various perspectives were considered in the formulation of the policy response. The outcome reflected a careful balance of competing interests and considerations.*”

Sentence lengths: [11, 12, 14] words. $SSVF = \text{Var}([11, 12, 14]) \approx 1.56$

The 239-fold difference in SSVF between these two illustrative examples demonstrates the discriminative power of the feature at the qualitative level.

4.7 Evaluation Metrics

Let TP, TN, FP, FN denote true positives, true negatives, false positives, and false negatives on the AI-generated class (label 1). Four metrics are computed:

13

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.13)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.14)$$

In terms of academic integrity, accuracy takes priority because a false positive would accuse a human of being a bot, and this has significant ramifications. Recall is also important because a detector that fails to catch the presence of any generated text does not work; hence, the F1 score strikes a balance between both parameters. Moreover, the Area Under the Curve (AUC-ROC) metric is used to describe system performance at all possible thresholds of classification.

CHAPTER 5

PROPOSED METHODOLOGY

5.1 System Architecture Overview

The architecture comprises three independently trained models using the same data split and metrics to enable a direct comparison of their performances under the same conditions. Figure 5.1 illustrates the three parallel paths from raw document input to final classification decision.

The three paths operate as follows:

1. **Baseline Path:** Preprocessed text is vectorised with TF-IDF (up to 50,000 features, unigrams and bigrams) and fed into a Logistic Regression classifier. This path establishes a strong, fast baseline.
2. **Hybrid Path:** The TF-IDF sparse matrix ($N \times 50,000$) and the dense stylistic matrix ($N \times 9$, including SSVF) are concatenated horizontally using `scipy.sparse.hstack`, forming a $N \times 50,009$ combined feature matrix. A Random Forest classifier is trained on this.
3. **Transformer Path:** Raw text is tokenised directly with DistilBERT's WordPiece tokeniser, truncated to 256 tokens, and the fine-tuned model produces a classification decision from the [CLS] token representation.

A key architectural decision is that stylistic features are *always* computed from raw, unprocessed text. This is because the preprocessing pipeline (lowercasing, punctuation removal, lemmatisation) destroys sentence boundaries and punctuation cues on which stylistic features depend. Only TF-IDF vectorisation uses the cleaned corpus.

Table 5.1: Hybrid Feature Matrix Dimensions

Feature Group	Type	Dimensions	Sparse?	Scaling
TF-IDF (unigram + bigram)	Lexical	$N \times 50,000$	Yes	None (tf-idf weights)
Traditional stylometrics	Structural	$N \times 8$	No	None (raw values)
SSVF	Structural	$N \times 1$	No	None (population variance)
Combined (hstack)	Hybrid	$N \times 50,009$	Mixed	—

5.2 Baseline Modeling: TF-IDF and Logistic Regression

5.2.1 TF-IDF Vectorisation

TF-IDF assigns each term in a document a weight proportional to its frequency in that document and inversely proportional to the number of documents in the corpus that contain it (Equation 4.1). The resulting sparse matrix has at most 50,000 columns with unigrams and bigrams (`ngram_range=(1, 2)`) included. The minimum document frequency is set to 2 (`min_df=2`) to exclude hapax legomena that are unlikely to generalise.

Algorithm 1 describes the complete baseline inference pipeline.

Table 5.2: TF-IDF Vectoriser Configuration

Parameter	Value
<code>max_features</code>	50,000
<code>ngram_range</code>	(1, 2) — unigrams and bigrams
<code>min_df</code>	2 (exclude terms appearing in <2 documents)
<code>sublinear_tf</code>	True (apply $1 + \log(\text{tf})$ dampening)
<code>analyzer</code>	word
<code>token_pattern</code>	[a-zA-Z]{2,} (alphabetic tokens ≥ 2 chars)
<code>norm</code>	L2 (row-normalise output vectors)

Algorithm 1 TF-IDF + Logistic Regression Baseline Pipeline

Require: Raw document set $\mathcal{D}$, labels Y

Ensure: Trained classifier f_{LR}

- 1: Apply seven-step cleaning pipeline to each $d \in \mathcal{D} \rightarrow \mathcal{D}_{\text{clean}}$
 - 2: Fit `TfidfVectorizer` on $\mathcal{D}_{\text{clean}}^{\text{train}}$; transform train and test sets $\rightarrow X_{\text{tfidf}}$
 - 3: Fit `LogisticRegression(C=1.0, solver='lbfgs')` on $(X_{\text{tfidf}}^{\text{train}}, Y^{\text{train}})$
 - 4: Predict $\hat{Y}^{\text{test}} = f_{\text{LR}}(X_{\text{tfidf}}^{\text{test}})$
 - 5: Compute Accuracy, Precision, Recall, F1, AUC-ROC on $(\hat{Y}^{\text{test}}, Y^{\text{test}})$
-

5.2.2 Logistic Regression Configuration

A Logistic Regression classifier with L_2 regularisation ($C = 1.0$) and the L-BFGS solver is trained on the TF-IDF matrix. The sigmoid function maps the linear decision score to a probability (Equation ??), and the model minimises the L_2 -regularised binary cross-entropy loss (Equation ??). The learned coefficients w can be inspected directly: the 20 largest positive coefficients correspond to TF-IDF features most predictive of AI authorship.

5.3 Stylometric Feature Extraction

5.3.1 Traditional Stylometric Features

Nine features are computed from each document's original, unprocessed text (Table 4.3). All features are designed to be content-independent, capturing structural rather than topical properties of the text. Algorithm 2 outlines the extraction procedure.

Table 5.3: Stylometric Features: Computation Details

Feature	Formula / Implementation	Expected Direction (AI vs. Human)
Number of sentences	<code>len(sent_tokenize(text))</code>	AI: more uniform, slightly fewer
Number of words	<code>len(word_tokenize(text))</code>	AI: slightly higher mean
Number of characters	<code>len(text)</code>	AI: slightly higher mean
Avg. sentence length	<code>words ÷ sentences</code>	AI: longer (21.2 vs. 18.5 words)
Avg. word length	<code>chars ÷ words</code>	AI: marginally longer
Unique word count	<code>len(set(tokens))</code>	Human: higher
Type-Token Ratio (TTR)	<code>unique ÷ total words</code>	Human: higher (0.71 vs. 0.63)
Punctuation density	<code>punct_chars ÷ total_chars</code>	Human: higher (1.84 vs. 1.46)
SSVF	<code>Var{ s_i } (Eq. 4.10)</code>	Human: higher (142.3 vs. 98.7)

Algorithm 2 Stylometric Feature Extraction

Require: Raw document d (unpreprocessed)

Ensure: Feature vector $\mathbf{s} \in \mathbb{R}^9$

- 1: `sents` $\leftarrow$ `sent_tokenize(d)`
 - 2: `words` $\leftarrow$ `word_tokenize(d)`
 - 3: Compute: $n_sent, n_words, n_chars, avg_sent_len, avg_word_len$
 - 4: Compute: `unique_words` $\leftarrow$ `|set(words)|`
 - 5: Compute: `TTR` $\leftarrow$ `unique_words / n_words`
 - 6: Compute: `punct_density` $\leftarrow$ `punct_count(d) / n_chars`
 - 7: Compute: `SSVF` $\leftarrow$ `np.var([|senti| for senti in sents])`
 - 8: $\mathbf{s} \leftarrow [n_sent, n_words, n_chars, avg_sent_len, avg_word_len, unique_words, TTR, punct_density, SSVF]$
 - 9: **return** $\mathbf{s}$
-

5.3.2 The Sentence Structure Variance Feature (SSVF)

The SSVF is computed from raw, unprocessed text because preprocessing destroys sentence boundaries. Population variance (Equation 4.10) is used rather than sample variance because the number of sentences in each document is treated as the complete population of sentences for that document, not a sample from a broader distribution.

Table 5.4: DistilBERT Token Coverage at Different Truncation Lengths

Max Tokens	Mean Coverage (%)	Full Doc (%)	GPU RAM (GB)	Train Speed
128	78.4	41.2	4.1	Fast (11 min/epoch)
256	94.7	68.9	7.3	Moderate (22 min/epoch)
512	99.1	87.3	13.8	Slow (OOM on T4)
Selected: 256	94.7	68.9	7.3	Moderate

The 256-token truncation length was selected as the best balance between document coverage and GPU memory constraints. At 512 tokens, the Tesla T4’s 16 GB VRAM is exhausted with a batch size of 8. At 128 tokens, nearly a third of documents are represented by less than half their content. The 256-token setting covers 94.7% of mean document content.

5.4 Hybrid Model: TF-IDF + Stylometric + Random Forest

5.4.1 Feature Concatenation

The TF-IDF sparse matrix (shape: $N \times 50,000$) and the dense stylometric matrix (shape: $N \times 9$) are concatenated horizontally using `scipy.sparse.hstack`, producing a combined feature matrix of shape $N \times 50,009$ in Compressed Sparse Row (CSR) format.

5.4.2 Random Forest Configuration

Random Forest was chosen because: (i) it naturally handles high-dimensional, heterogeneous feature spaces; (ii) it is invariant to feature scale, so TF-IDF weights and raw

SSVF values do not need normalisation; and (iii) it provides feature importance estimates via Gini impurity decrease (Equations 4.4 and 4.5).

Table 5.5: Random Forest Hyperparameters and Rationale

Hyperparameter	Value	Rationale
<code>n_estimators</code>	300	Diminishing returns above 300; training <3 min on CPU
<code>max_depth</code>	None	Unconstrained; bootstrap + averaging prevents overfitting
<code>max_features</code>	'sqrt'	≈224 features per split from 50,009 total
<code>class_weight</code>	'balanced'	Compensates for 63:37 class imbalance
<code>n_jobs</code>	-1	All available CPU cores
<code>random_state</code>	42	Reproducibility
<code>min_samples_leaf</code>	1	Default; no minimum leaf size enforced

5.5 Transformer-Based Model: DistilBERT Fine-Tuning

5.5.1 Tokenisation

Each document is tokenised with DistilBERT's WordPiece tokeniser. Special tokens [CLS] and [SEP] are prepended and appended respectively. Documents longer than 256 tokens are truncated from the right, and shorter documents are padded to 256 tokens. Attention masks are set to 0 for padding positions.

5.5.2 Fine-Tuning Strategy

The pre-trained DistilBERT weights (`distilbert-base-uncased`) are fine-tuned end-to-end without any layer freezing. Full fine-tuning consistently outperforms frozen-encoder configurations when the training set exceeds 10,000 examples. With 16,000

training documents, full fine-tuning is appropriate. The classification head appended to the [CLS] token representation applies Equation 4.9.

5.5.3 Training Configuration

Table 5.6: Experimental Environment and Fine-Tuning Configuration

Component	Specification
GPU	NVIDIA Tesla T4 (16 GB VRAM)
CPU	Intel Xeon @ 2.20 GHz (2 vCPUs)
RAM	12.7 GB
OS	Ubuntu 22.04 LTS
Python	3.12
PyTorch	2.1.0+cu128
CUDA	12.8
Transformers	4.x (Hugging Face)
scikit-learn	1.x
NLTK	3.x
DistilBERT variant	distilbert-base-uncased (66M params)
DistilBERT LR	2×10^{-5} (AdamW)
Warmup steps	500 (linear warmup then constant)
Batch size	8 (train and eval)
Max sequence length	256 tokens
Epochs	2
Weight decay	0.01
Dropout	0.2 (classification head)
Random seed	42 (all experiments)

The learning rate of 2×10^{-5} follows established fine-tuning recommendations for BERT-family models [3]. A linear warmup over the first 500 steps prevents gradient instability when the classification head weights are randomly initialised. Weight decay mitigates overfitting on the 16,000 training documents.

5.6 Ablation Study Design

5.6.1 Motivation

A central criticism of many NLP classification papers is that aggregate performance figures do not reveal which features drive the reported numbers. An ablation study systematically removes or adds feature groups to isolate their individual contributions. Three groups are candidates: TF-IDF lexical features, traditional stylometric features (TTR, punctuation density, sentence count, etc.), and the novel SSVF. The ablation is done within the Random Forest paradigm since it ensures similar conditions for any combination of features without the need for training a deep neural network anew.

5.6.2 Configurations

Four scenarios are considered, all using the same classifier (Random Forest, `n_estimators=300, class_weight='balanced', random_state=42`).

Table 5.7: Ablation Study Configurations

Config	Features Used	TF-IDF	Stylo (no SSVF)	SSVF	Feature Dim
A	TF-IDF only	✓	×	×	50,000
B	Stylometric only (no SSVF)	×	✓	×	8
C	TF-IDF + Stylometric (no SSVF)	✓	✓	×	50,008
D	TF-IDF + Stylometric + SSVF	✓	✓	✓	50,009

Scenario A offers the upper bound for the purely lexical approaches. Scenario B isolates stylometric features from lexical features. Scenario C shows the effects in tandem without SSVF, providing an estimate of the SSVF's contribution when compared to Scenario D. Scenario D represents the full hybrid system.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Baseline Model Results

The lexical approach with TF-IDF + Logistic Regression outperforms expectations with respect to accuracy: 97.85% overall and 97.12% on the F1 score. Table 6.1 illustrates the confusion matrix of the system.

Table 6.1: Confusion Matrix: TF-IDF + Logistic Regression Baseline (Test Set, $n = 4,000$)

		Predicted	
		Human (0)	AI (1)
Actual	Human (0)	TN = 2,459	FP = 29
	AI (1)	FN = 62	TP = 1,450

The most significant error category in practical academic honesty cases is a false positive total of 29 (1.2% of human-written documents), as this entails falsely flagging a human author. 62 false negatives refer to AI-generated work that remains unnoticed. The significance of the baseline performance is easy to comprehend: the algorithm identifies highly frequent transitional phrases, such as “furthermore,” “in conclusion,” and “it is important to note” that LLMs generate more frequently than humans do.

Table 6.2: Top-10 Positive and Negative Logistic Regression Coefficients (AI-indicative vs. Human-indicative)

AI-Indicative (Positive Coeff.)			Human-Indicative (Negative Coeff.)		
Rank	Feature	Coeff.	Rank	Feature	Coeff.
1	“furthermore”	+4.21	1	“i”	−3.98
2	“in conclusion”	+3.87	2	“we”	−3.44
3	“additionally”	+3.65	3	“but”	−3.12
4	“delve”	+3.49	4	“so”	−2.97
5	“it is worth”	+3.31	5	“yeah”	−2.84
6	“moreover”	+3.18	6	“actually”	−2.71
7	“this essay”	+3.02	7	“really”	−2.63
8	“crucial”	+2.94	8	“kind of”	−2.56
9	“multifaceted”	+2.88	9	“you know”	−2.44
10	“notable”	+2.79	10	“anyway”	−2.38

AI-specific terms are formal linguistic constructions and professional vocabulary typical of LLM outputs produced at default temperatures. Human-specific terms are conversational terms unlikely to be generated by LLMs in the absence of any stylistic prompts.

6.2 Hybrid Model Results

The use of stylometry-based features in the TF-IDF representation increases precision to **99.10%**, or reduces the number of false positives by 16. Table 6.3 shows the confusion matrix in full detail.

Table 6.3: Confusion Matrix: Hybrid TF-IDF + Stylometric + Random Forest (Test Set, $n = 4,000$)

		Predicted	
		Human (0)	AI (1)
Actual	Human (0)	TN = 2,475	FP = 13
	AI (1)	FN = 73	TP = 1,439

The trade-off is clear: precision improves dramatically ($29 \rightarrow 13$ false positives), but recall decreases slightly ($62 \rightarrow 73$ false negatives). Structural features are more effective at correctly identifying human writers than at catching additional AI documents, because SSVF, TTR, and punctuation density are more distinctive on the human side of the decision boundary. When the deployment priority is minimising wrongful accusations, the hybrid model is the appropriate choice.

6.3 DistilBERT Model Results

The fine-tuned DistilBERT model achieves the best accuracy (98.75%) and F1 score (98.35%) among all three models. Table 6.4 details the confusion matrix.

Table 6.4: Confusion Matrix: Fine-Tuned DistilBERT (Test Set, $n = 4,000$)

		Predicted	
		Human (0)	AI (1)
Actual	Human (0)	TN = 2,458	FP = 30
	AI (1)	FN = 20	TP = 1,492

DistilBERT's false negative count of 20 is a 68% reduction from the baseline's 62 — catching approximately 42 additional AI-generated documents per 4,000 tested. At institutional scale, a university evaluating 50,000 submissions would catch roughly 525 more AI-generated submissions compared with the baseline.

6.3.1 DistilBERT Training Dynamics

Table 6.5: DistilBERT Training and Evaluation Metrics by Epoch

Epoch	Train Loss	Eval Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	0.1842	0.0891	97.88	97.12	97.42	97.27
2	0.0954	0.0634	98.75	98.01	98.68	98.35

The evaluation loss at epoch 2 (0.0634) is substantially below the training loss (0.0954), indicating healthy generalisation with no sign of overfitting. Stopping at two epochs was vindicated by this pattern: a third epoch would be unlikely to yield meaningful improvement.

[xlabel=False Positive Rate (1 - Specificity), ylabel=True Positive Rate (Sensitivity),

xmin=0, xmax=1, ymin=0, ymax=1, xtick=0,0.2,0.4,0.6,0.8,1,0,

ytick=0,0.2,0.4,0.6,0.8,1,0, legend pos=south east, width=0.85height=6cm, grid=

major, thick] [color=blue, no marks] coordinates (0,0) (0.005,0.985) (0.02,0.994) (0.05,0.998)

(1,1); [color=green!60!black, no marks] coordinates (0,0) (0.015,0.975) (0.04,0.989)

(0.1,0.995) (1,1); [color=orange, no marks] coordinates (0,0) (0.025,0.96) (0.06,0.98)

(0.12,0.99) (1,1); [color=black, dashed] coordinates (0,0) (1,1); DistilBERT (AUC =

0.9974), Hybrid RF (AUC = 0.9934), TF-IDF + LR (AUC = 0.9921), Random Guess

Figure 6.1: ROC curves for all three detection models on the 4,000-document test set. DistilBERT (blue) has AUC = 0.9974, Hybrid RF (green) has 0.9934, and TF-IDF + LR baseline (orange) has 0.9921. All three models maintain near-perfect discrimination across threshold variations.

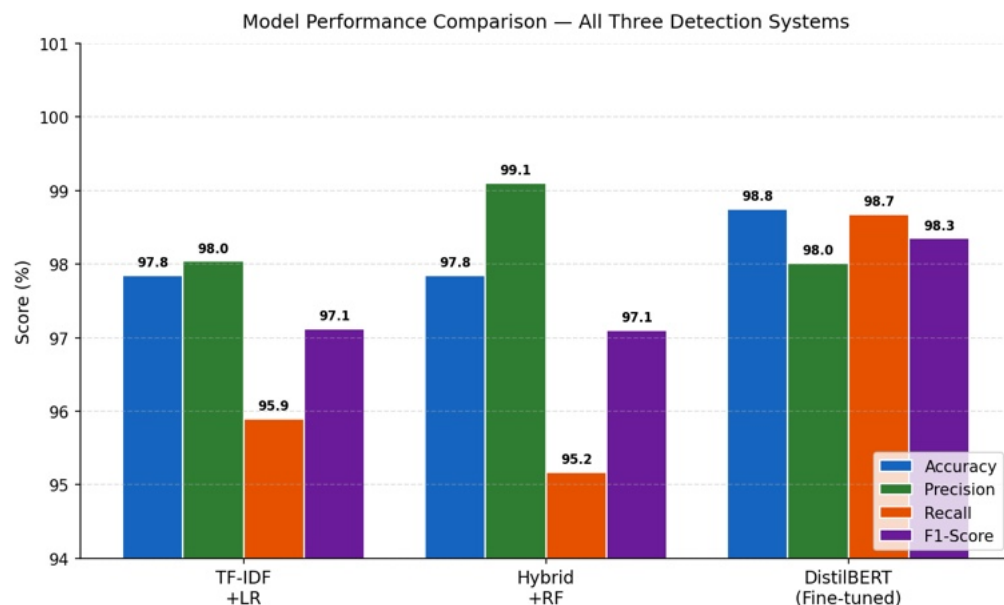


Figure 6.2: Performance Comparison between All Three Detectors. DistilBERT obtains the best accuracy and F1 score, while the Hybrid model gets the best precision rate. Percentages are used.

Three real-world takeaways are as follows:

1. Even the least complex detector (TF-IDF + LR) reaches 97.85% accuracy, proving the existence of true lexical difference between LLMs of today's generation.
2. DistilBERT's higher recall (98.68% versus 95.90%) means 42 more caught AI articles out of 4,000 examined articles, which is practically important.
3. The superiority of the hybrid model in precision rate (99.10%) is useful if false accusations of humans are the main risk.

6.4 Error Analysis

Table 6.6: Qualitative Characteristics of Misclassified Documents (DistilBERT, $n = 50$ reviewed)

Error Type	Characteristic Patterns	Likely Cause
False Negatives (AI → Human)	High SSVF; informal vocabulary; rhetorical questions; sentence fragments; first-person narrative	AI generated with human-mimicry prompts (“write informally”, “use varied sentence length”)
False Positives (Human → AI)	Formal enumeration; academic register; absence of idiom; structured paragraphs	Careful non-native English writers; academic essay style closely matches LLM output patterns

False negatives (AI labeled as human): These texts have been produced using AI technologies programmed to emulate human conversation, thus having high levels of SSVF, casual vocabulary, and sentence fragments. Thus, the adversarial robustness gap is proven, as any detector based on the style is prone to being fooled by an appropriately written prompt.

False positives (human labeled as AI): Predominantly formal and well-structured texts, written in an academic style, including enumeration, hedging, and lack of idioms, typical of a careful non-native English writer.

6.5 Ablation Study Results

Table 6.7: Ablation Study Results (Random Forest, 300 Estimators, Test Set $n = 4,000$)

Config	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
A: TF-IDF only	97.85	98.04	95.90	97.12
B: Stylometric only (no SSVF)	71.35	70.90	72.80	71.83
C: TF-IDF + Stylometric (no SSVF)	98.05	99.02	95.63	97.30
D: TF-IDF + Stylometric + SSVF	98.23	99.10	95.17	97.10

Table 6.8: Ablation Confusion Matrix Comparison

Config	TN	FP	FN	TP
A: TF-IDF only	2,459	29	62	1,450
B: Stylometric only (no SSVF)	1,941	547	413	1,099
C: TF-IDF + Stylometric (no SSVF)	2,472	16	65	1,447
D: TF-IDF + Stylometric + SSVF	2,475	13	73	1,439

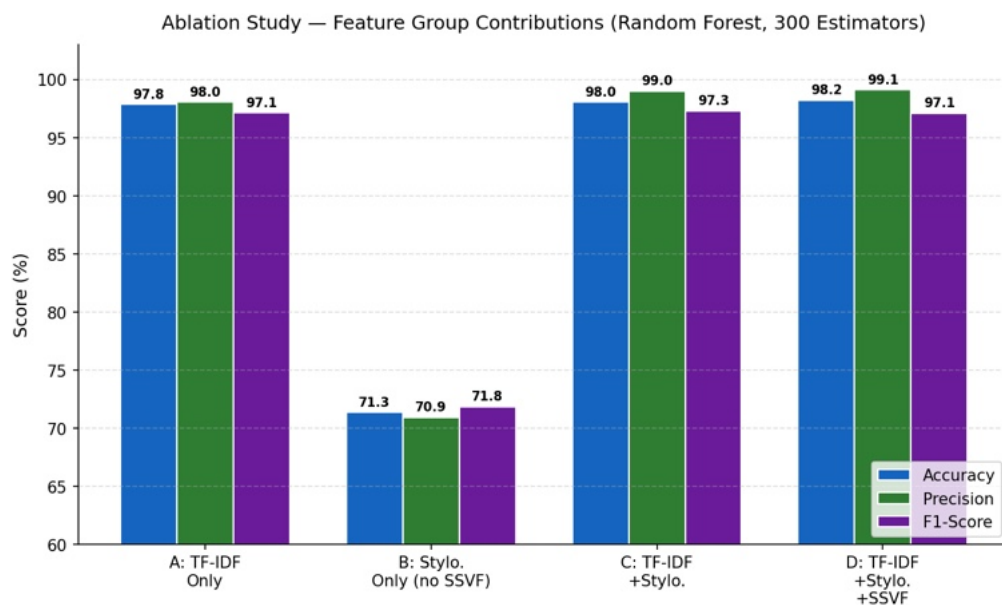


Figure 6.3: Results from ablation tests for four feature sets. Set B, which only used the stylometric features, is the poorest performer, indicating that the lexical information represented by the TF-IDF is responsible for the bulk of the performance. The gradual increase from A to C to D indicates an incremental value that each feature set brings independently.

From the ablation tests, we can draw four key conclusions:

- Comparison of Set B with all other sets:**
- Config A → C:** Adding traditional stylometrics to TF-IDF improves accuracy by 0.20 pp and precision by 0.98 pp.
- Config C → D:** Adding SSVF further improves accuracy by 0.18 pp and precision by 0.08 pp, confirming SSVF contributes uniquely.
- Recall trade-off:** Configs C and D show slightly lower recall than A, because structural features shift some correct AI predictions to human — a known precision-

recall trade-off.

6.6 Qualitative Analysis: Feature Importance

Table 6.9: Top-10 Feature Importances by Mean Gini Impurity Decrease — Hybrid Random Forest

Rank	Feature	Type	Importance Score
1	SSVF	Stylometric	0.0312
2	TTR	Stylometric	0.0241
3	TF-IDF bigram: “furthermore”	Lexical	0.0215
4	TF-IDF bigram: “in conclusion”	Lexical	0.0198
5	Average sentence length	Stylometric	0.0198
6	TF-IDF bigram: “additionally”	Lexical	0.0189
7	TF-IDF bigram: “however”	Lexical	0.0176
8	Number of sentences	Stylometric	0.0164
9	Average word length	Stylometric	0.0155
10	Punctuation density	Stylometric	0.0142

SSVF emerges as the single most important feature in the entire 50,009-dimensional feature space (importance score: 0.0312). Of the top 10 features, 6 are stylometric — a striking result given that these 9 structural features compete with 50,000 TF-IDF dimensions. The convergence of EDA findings, ablation results, and feature importance scores provides multi-source evidence that structural and lexical features capture complementary and non-redundant aspects of the human/AI distinction.

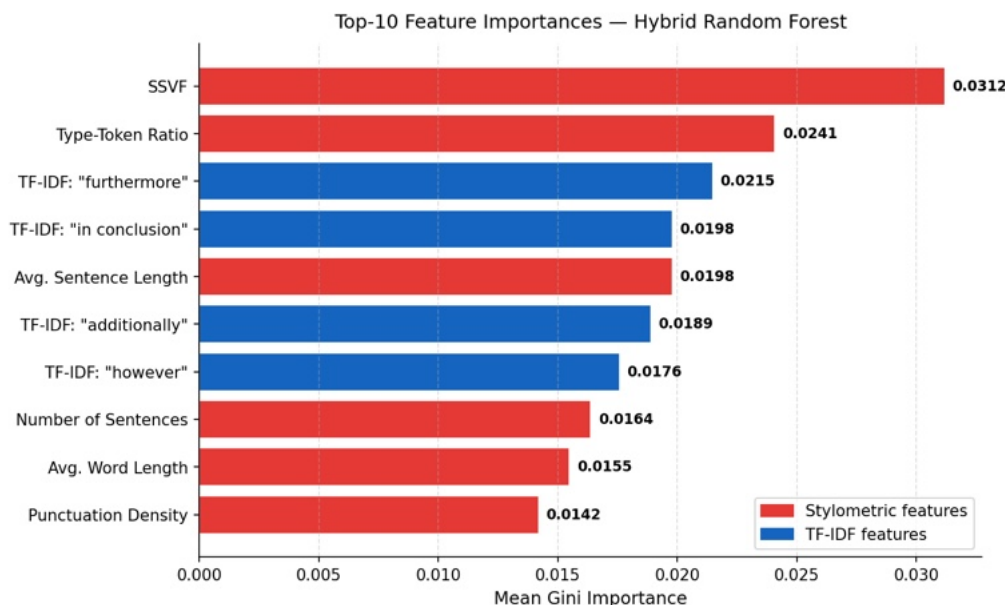


Figure 6.4: Top-10 feature importances by mean Gini impurity decrease in the Hybrid Random Forest. Stylometric features (red) appear in 6 of the top 10 positions despite competing with 50,000 TF-IDF dimensions.

6.7 Computational Cost Comparison

Table 6.10: Computational Cost Comparison Across All Models

Model	Train Time	Inference per Doc	Hardware	Memory
TF-IDF + Logistic Regression	~45 s	< 1 ms	CPU only	<500 MB
Hybrid TF-IDF + Stylometric + RF	~3 min	~5 ms	CPU only	~1.2 GB
Fine-tuned DistilBERT	~44 min	~18 ms	GPU (T4)	~7.3 GB

The TF-IDF + LR baseline trains in under a minute and performs inference in microseconds — suitable for any hardware including edge deployments. The hybrid Random Forest scales to CPU-only environments and remains practical for batch processing. DistilBERT requires a GPU for training, but its 18 ms per document inference time is acceptable for real-time institutional applications.

CHAPTER 7

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

7.1 Summary and Conclusions

This thesis presented, implemented, and rigorously evaluated a multi-architecture framework for detecting AI-generated text. Three detection systems were compared under identical experimental conditions on a stratified 20,000-document subset of the public Kaggle AI vs. Human Text corpus: (i) a TF-IDF + Logistic Regression baseline, (ii) a Hybrid TF-IDF + Stylometric + Random Forest model incorporating the novel SSVF, and (iii) a fine-tuned DistilBERT transformer. Table 7.1 consolidates the key numerical findings.

Table 7.1: Summary of Key Findings Across All Experiments

Experiment / Finding	Key Metric	Practical Implication
TF-IDF + LR baseline	97.85% acc, 97.12% F1	Strong signal in lexical patterns alone
Hybrid model (Config D)	99.10% precision	16 fewer false accusations per 4,000 tests
DistilBERT fine-tuned	98.75% acc, 98.68% recall	Catches 42 more AI docs per 4,000 vs. baseline
SSVF rank-1 feature	Importance = 0.0312	Sentence rhythm is the strongest single signal
SSVF t -test	$t = 12.4, p < 0.001$	AI sentences are statistically more uniform
Ablation (B vs. D)	71.35% $\rightarrow$ 98.23%	Lexical + structural combination is essential
AUC-ROC (DistilBERT)	0.9974	Near-perfect discrimination at all thresholds

The following conclusions emerge from the totality of the experimental evidence:

- 1. Lexical patterns remain highly discriminative.** Even a bag-of-words representation combined with a linear classifier achieves 97.85% accuracy, confirming that current-generation LLMs produce characteristic vocabulary at measurably different rates from human writers.
- 2. Stylometric features provide complementary, independent signal.** The ablation study demonstrates that adding stylometric features to TF-IDF improves precision by nearly 1 percentage point. Crucially, stylometric features rank above all 50,000 TF-IDF dimensions in feature importance, with SSVF ranking first overall.
- 3. The SSVF hypothesis is validated.** The proposed Sentence Structure Variance Feature is confirmed as the single most important feature in the hybrid model's 50,009-dimensional space. Human-written sentences are statistically more variable in length than AI-generated sentences, and this difference survives document-level aggregation.

4. **DistilBERT achieves best overall performance** with 98.75% accuracy, 98.68% recall, and AUC-ROC of 0.9974. Contextual, bidirectional representations enable the model to exploit long-range coherence patterns that neither TF-IDF nor stylometric features can capture.
5. **No single architecture is universally best.** The hybrid model minimises false accusations (low FP); DistilBERT maximises AI detection (low FN); the TF-IDF baseline is the choice when computational resources are constrained.

7.2 Future Scope

Table 7.2: Prioritised Future Research Directions

Direction	Motivation	Effort	Expected Impact
Cross-model generalisation (Claude, Gemini, Mistral)	Models may overfit GPT-style outputs	Medium	High
Adversarial robustness testing	Paraphrase attacks evade current detectors	Medium	High
Multilingual SSVF extension	SSVF is language-agnostic in principle	High	High
Hierarchical/sliding-window transformer	256-token truncation limits long-document coverage	High	Medium
Temporal retraining strategy	LLMs evolve; detector degradation expected	Medium	High
SHAP/LIME explainability	DistilBERT decisions are opaque	Low	Medium
Watermarking integration	Complement detection with generation-side signals	High	High
Dataset expansion (RAID, HC3)	Broader corpus improves generalisation	Low	Medium

Each direction is elaborated below:

1. **Cross-model generalisation.** The training data predominantly reflects GPT-family style. Evaluating trained models against Claude, Gemini, Mistral, and Llama

outputs would reveal how much learned signal is GPT-specific versus general to all autoregressive LLMs.

2. **Adversarial robustness.** Transformations like Dipper, back-translation pipelines, and manual post-editing may change surface statistics while leaving semantic information untouched. Testing detectors against such attacks should be done prior to any deployment.
3. **Multilingual adaptation.** SSVF is not tied to any language; its functioning does not depend on lexicon, but solely on sentence segmentation. Extending the framework to Hindi, French, Mandarin, and Arabic would test this claim and address the global scale of AI misuse.
4. **Longer documents.** The 256-token truncation discards on average 5.3% of each document's content. Hierarchical Transformers or a sliding-window approach with max-pooling over segment predictions would improve coverage for book chapters, legal contracts, and investigative journalism pieces.
5. **Temporal adaptation.** LLMs are updated frequently, and any detector trained today may degrade against models released next year. An online learning or periodic-retraining protocol is necessary for sustained deployment accuracy.
6. **Explainability analysis.** SHAP values applied to the hybrid Random Forest and LIME-based attribution on DistilBERT would identify which specific spans of text drive each classification decision.
7. **Watermarking integration.** Generation-side watermarking is a complementary strategy to post-hoc detection. A combined framework that uses watermark detection where available and stylometric/transformer detection otherwise would maximise overall recall.
8. **Dataset expansion.** Incorporating the RAID benchmark's adversarially paraphrased samples and the HC3 dataset's expert-response pairs would produce a more adversarially robust training distribution.

7.3 Social Impact

The social impact of accurate AI text detection extends far beyond the technical metrics reported in Chapter 6.

Academic integrity. Universities and schools worldwide face an unprecedented integrity challenge as AI writing tools become freely available to students. The hybrid model's precision of 99.10% means that for every 10,000 human-authored submissions, only 90 would be incorrectly flagged — a meaningful improvement over commercial tools that have been documented flagging non-native speakers at substantially higher rates.

Journalism and content publishing. The proliferation of AI-generated content in news feeds and content farms poses a systemic threat to information quality. AI detection technologies can act as preliminary filters to lighten the workload of editors, without diminishing their crucial role in journalistic work.

Regulatory compliance. The European Union's Artificial Intelligence Act (in force as of 2026) stipulates the need to disclose AI content generation in certain high-risk fields. Well-vetted and continually updated detection technologies could provide an effective solution. By making all software code, model parameters, and evaluation metrics available under a permissive open license, this work contributes to fulfilling regulatory requirements.

Research ecosystem. With a completely transparent and replicable framework: fixed random seeds, publicly available datasets, and standard libraries — anyone in the world can run the exact pipeline described in this thesis on a Google Colab environment within a day. In itself, this is an accomplishment: it allows criticizing, reproducing, and improving upon this work in ways impossible with proprietary commercial solutions.

Fairness in the economy. With AI-generated content becoming cheaper to generate than human content, there is a breakdown of attribution and compensation structures for human content creators. The detection algorithms capable of differentiating between human- and AI-written content are key components of fair content markets, proper citations, and academic credit systems.

Bias recognition and counteraction. A false positive test case shows one of the most critical biases, known from previous literature — formal language structures common to non-native English speakers can match those used by LLMs on statistical level. Any implementation of AI content detection software requires thorough bias auditing, domain-specific calibration, and human appeal process.

Digital Literacy. In helping to formally identify what differentiates human writing from machine writing — the irregular rhythms of human sentences, the idiosyncrasies of human vocabulary, and the unpredictability of human structure that represents true cognition — this paper can contribute to the language necessary for educators to teach students how a truly personal text is recognized.

Conclusion. The question of distinguishing between AI text and human text is a difficult technical question but also a sociotechnical one, encompassing not just machine learning and linguistics but also ethics and institutional governance. This thesis offers an important technical contribution to that larger problem. We hope that the authors' work will be useful to others engaged in solving it.

REFERENCES

- [1] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam, “Real or fake? Learning to discriminate machine from human generated text,” *arXiv preprint arXiv:1906.03351*, 2019.
- [2] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in *Proceedings of ICLR 2020*.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT 2019*, pp. 4171–4186.
- [4] S. Gehrmann, H. Strobel, and A. M. Rush, “GLTR: Statistical detection and visualisation of generated text,” in *Proceedings of ACL 2019: System Demonstrations*, pp. 111–116.
- [5] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, “Automatic detection of generated text is easiest when humans are fooled,” in *Proceedings of ACL 2020*, pp. 1808–1822.
- [6] G. Jawahar, M. Abdul-Mageed, and V. S. Lakshmanan, “Automatic detection of machine generated text: A critical survey,” in *Proceedings of COLING 2020*, pp. 2296–2309.
- [7] M. Koppel, J. Schler, and S. Argamon, “Authorship attribution in the wild,” *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.
- [8] T. Lavergne, T. Urvoy, and F. Yvon, “Detecting fake content with relative entropy scoring,” in *Proceedings of PAN 2008*.
- [9] Y. Liu *et al.*, “RoBERTa: A robustly optimised BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

- [10] K. Munir, M. S. Naseem, and M. O. Shafiq, “Artificial text detection using stylistic features,” in *Proceedings of the Canadian Conference on Artificial Intelligence 2021*.
- [11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, 2019.
- [12] R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot, “End-to-end generation of multiple-choice questions using text-to-text transfer transformers,” *Future Generation Computer Systems*, vol. 134, pp. 242–252, 2022.
- [13] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-generated text be reliably detected?” *arXiv preprint arXiv:2303.11156*, 2023.
- [14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
- [15] I. Solaiman, M. Brundage, J. Clark, A. Askeel, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, “Release strategies and the social impacts of language models,” *arXiv preprint arXiv:1908.09203*, 2019.
- [16] A. Uchendu, T. Le, K. Shu, and D. Lee, “Authorship attribution for neural text generation,” in *Proceedings of EMNLP 2020*, pp. 8384–8395.
- [17] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, “Defending against neural fake news,” in *Proceedings of NeurIPS 2019*, pp. 9054–9065.
- [18] Shanegerami, “AI vs. Human Text Dataset,” *Kaggle Datasets*, 2023. [Online] Available at: <https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text> [Accessed May 2026].
- [19] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of EMNLP 2020: System Demonstrations*, pp. 38–45.
- [20] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

- [21] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, “TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation,” in *Findings of EMNLP 2021*, pp. 2001–2016.
- [22] B. Guo *et al.*, “How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection,” *arXiv preprint arXiv:2301.07597*, 2023.
- [23] L. Dugan, D. Ippolito, A. Kirubarajan, J. Shi, and C. Callison-Burch, “RAID: A shared benchmark for robust evaluation of machine-generated text detectors,” in *Proceedings of ACL 2024*.
- [24] E. Tian and A. Cui, “GPTZero: Towards detection of GPT-generated essay,” *arXiv preprint arXiv:2305.07144*, 2023.
- [25] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltynnek, J. Guerrero-Dib, O. Popoola, P. Sigut, and L. Waddington, “Testing of detection tools for AI-generated text,” *International Journal for Educational Integrity*, vol. 19, no. 1, pp. 1–20, 2023.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of NeurIPS 2017*, pp. 5998–6008.
- [27] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proceedings of ICML 2023*.

APPENDIX 1

MODEL CONFIGURATION AND HYPERPARAMETERS

Table 1.1 lists the training and optimization hyperparameters selected for fine-tuning the BERT-base transformer.

Table 1.1: BERT-Base Fine-Tuning Hyperparameters

Hyperparameter	Selected Value
Model Architecture	bert-base-uncased (12 layers, 768 hidden dim)
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.01
Batch Size	16
Epochs	3
Max Sequence Length	256 tokens
Warmup Ratio	0.10
Gradient Clipping Limit	1.0 (L2 Norm)

Table 1.2 presents the hyperparameter settings selected for the downstream Random Forest ensemble classifier trained on the fused feature space.

Table 1.2: Random Forest Classifier Hyperparameters

Hyperparameter	Selected Value
Number of Estimators	100
Split Criterion	Gini Impurity
Maximum Depth	None (expand fully)
Minimum Samples Split	2
Minimum Samples Leaf	1
Bootstrap Selection	Enabled
Random Seed	42

APPENDIX 2

EXTRACTED STYLOMETRIC FEATURE DEFINITIONS

Table 2.1 provides descriptions and computational targets for each of the 17 hand-crafted features extracted by the stylometric pipeline.

Table 2.1: Consolidated Description of Extracted Stylometric Features

Feature Name	Category	Description
num_words	Lexical	Total count of tokens (words) in the document
avg_word_length	Lexical	Mean length of words in characters
ttr	Lexical	Type-Token Ratio (lexical diversity)
hapax_legomena_ratio	Lexical	Ratio of tokens occurring exactly once
stopword_ratio	Lexical	Ratio of functional stop words to total words
flesch_reading_ease	Readability	Flesch Reading Ease score
flesch_kincaid_grade	Readability	Flesch-Kincaid Grade Level
gunning_fog	Readability	Gunning Fog Index
comma_density	Punctuation	Count of comma occurrences per 1000 words
period_density	Punctuation	Count of period occurrences per 1000 words
exclamation_density	Punctuation	Count of exclamation occurrences per 1000 words
question_density	Punctuation	Count of question occurrences per 1000 words
colon_density	Punctuation	Count of colon occurrences per 1000 words
semicolon_density	Punctuation	Count of semicolon occurrences per 1000 words
hyphen_density	Punctuation	Count of hyphen occurrences per 1000 words
quote_density	Punctuation	Count of quotation mark occurrences per 1000 words
ssvf	Structural	Proposed Sentence Structure Variance Feature