# Human vs. AI Text Detection: A Comparative Study Using Traditional and Deep Learning Models

## Abstract

This paper investigates the application of various machine learning and deep learning architectures for the binary classification of text as either human-written or AI-generated. Using a comprehensive dataset of text samples, we implemented and evaluated a suite of models including Logistic Regression, Support Vector Machines, LSTMs, and fine-tuned BERT. The BERT model demonstrated superior performance, achieving 98% accuracy, highlighting the power of transformer-based models in discerning subtle stylistic differences between human and AI authors.

## Introduction

The proliferation of sophisticated large language models (LLMs) has made distinguishing human-written text from AI-generated content a critical challenge in academia, journalism, and cybersecurity. This task is essential for ensuring authenticity and combating misinformation. This study explores the efficacy of different classification paradigms, from traditional machine learning on TF-IDF features to advanced deep learning sequence models. By comparing model performance, we aim to identify the most effective approach for this emerging and important natural language processing (NLP) task.

## Data Preprocessing

The dataset consisted of a large collection of text samples labeled as 'human' or 'AI-generated'. To prepare the text for modeling, a standard NLP preprocessing pipeline was applied:

* Cleaning: Removal of non-alphanumeric characters and digits.
* Normalization: Conversion to lowercase and stripping of extra whitespace.
* Stopword Removal: Common English stopwords were filtered out.
* Lemmatization: Words were reduced to their base or dictionary form using the WordNetLemmatizer.

## Model Building

We implemented a diverse set of models to tackle the classification problem:

1. Traditional Machine Learning (on TF-IDF Features):

- Logistic Regression: A linear model for binary classification.

- Support Vector Machine (SVM): A kernel-based model effective in high-dimensional spaces.

1. Neural Network (LSTM):

- A sequential model using an Embedding layer followed by two LSTM layers to capture sequential dependencies in the text.

1. Transformer-Based (BERT):

- The 'bert-base-uncased' model was fine-tuned for sequence classification with a maximum sequence length of 100 tokens.

1. Anomaly Detection (One-Class SVM with BERT Embeddings):

- BERT was used as a feature extractor, training a One-Class SVM on embeddings from human text.

## Training

Traditional models were trained on TF-IDF transformed text, LSTM was trained on 250,000 samples for 5 epochs, and BERT fine-tuned on 300,000 samples for 4 epochs. AdamW optimizer with a learning rate of 2e-5 and linear scheduler ensured stable convergence.

## Model Evaluation

Models were evaluated on held-out test sets. The BERT model significantly outperformed all other approaches.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision (AI) | Recall (AI) | F1-Score (AI) |
| Logistic Regression | 95% | 95% | 95% | 0.95 |
| SVM | 95% | 95% | 95% | 0.95 |
| LSTM Neural Network | 97% | 97% | 97% | 0.97 |
| Fine-Tuned BERT | 98% | 98% | 98% | 0.98 |
| One-Class SVM (BERT) | 93% | 92% | 95% | 0.93 |

## Conclusion

This study successfully demonstrates the capability of modern NLP techniques to distinguish between human and AI-generated text. Through rigorous preprocessing and model experimentation, we found that the fine-tuned BERT model achieved the highest performance with 98% accuracy, establishing it as the superior method for this task.

Deep learning models (LSTM, BERT) consistently outperformed traditional machine learning models, underscoring the importance of capturing complex, sequential patterns in text. Future work could focus on improving model interpretability, cross-domain generalization, and real-time detection capabilities.